Optimal shrinkage for robust covariance matrix estimators in a small sample size setting

Karina Ashurbekova, Antoine Usseglio-Carleve, Florence Forbes, Sophie Achard

Optimal shrinkage for robust covariance matrix estimators in a small sample size setting

Karina Ashurbekova, Antoine Usseglio-Carleve, Florence Forbes*, Sophie Achard

Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
firstname.lastname@inria.fr

Abstract

When estimating covariance matrices, traditional sample covariance-based estimators are straightforward but suffer from two main issues: 1) a lack of robustness, which occurs as soon as the samples do not come from a Gaussian distribution or are contaminated with outliers and 2) a lack of data, which occurs as soon as the covariance matrix dimension is greater than the sample size. The first issue can be handled by assuming that samples are drawn from a heavy-tailed distribution, at the cost of more complex derivations, while the second issue can be addressed by shrinkage with the difficulty of choosing the appropriate level of regularization. This work offers both a tractable and optimal framework based on shrinked likelihood-based M-estimators. First, a closed-form expression is provided for a regularized covariance matrix estimator with an optimal shrinkage coefficient for any sample distribution in the elliptical family. Then, a complete inference procedure is proposed which can handle both unknown mean and tail parameter, in contrast to most existing methods that focus on the covariance matrix parameter requiring pre-set values for the others. An illustration on synthetic and real brain connectivity data is provided in the case of the t-distribution with unknown mean and degrees-of-freedom parameters.

Keywords: Covariance estimation, Small sample size, Shrinkage methods,

*Corresponding author

October 2020

1. Introduction

Accurate estimation of covariance matrices (or more generally scale matrices) is fundamental in many areas of statistics and machine learning. Examples include applications in finance [1], bioinformatics and classification [2, 3]. Practitioners usually have to deal with two main difficulties. First, observations may deviate from the Gaussian distribution due to a particular data generating process or the presence of outlying data. Ignoring this deviation may conduct to inadequate predictions and conclusions [3]. A widespread solution to design so-called robust estimators, is to consider heavy-tailed distributions which can better accommodate outliers. Among those, elliptical distributions have been studied as good candidates as they include tractable heavy-tailed distributions such as the t-distribution, whose tail is controlled by a single degrees-of-freedom (d.o.f.) parameter [4, 5]. In addition, for elliptical distributions, robust estimators of the scale matrix $\mathbf{\Sigma}$ are provided by Maronna's M-estimators, defined as the solution $\widetilde{\mathbf{\Sigma}}$ of a fixed-point relationship $\widetilde{\mathbf{\Sigma}} = \mathbb{E}\left[u\left(\boldsymbol{x}^\top \widetilde{\mathbf{\Sigma}}^{-1}\boldsymbol{x}\right)\boldsymbol{x}\boldsymbol{x}^\top\right]$ where u is function satisfying a set of general assumptions [6]. Other robust approaches include the use of contamination models, see *e.g.* [7] for a recent reference. A second difficulty is then that the problem dimension may be too large compared to the number of available observations, which prevents accurate estimation when this feature is not explicitly taken into account. For example, if the dimension p of $\mathbf{\Sigma}$ is greater than the sample size n, Maronna's estimators do not exist [8]. As a consequence, many authors have proposed alternative estimators which can be divided into two main categories. A first set of approaches assumes structured matrices so as to reduce the number of parameters to estimate, while a second set of approaches aims at compensating the lack of samples with regularization or prior knowledge modelling. The first category includes attempts based on sparsity assumptions such as graphical Lasso, *e.g.* [9, 10, 11], and nodewise Lasso, *e.g.* [12, 13]. Besides not to be always satisfying in small

sample size settings (see [14] for a recent review), these methods assume Gaussian observations and are therefore not suitable for elliptical distributions with heavy tails. Generalizations and alternatives have been considered that are more robust, *e.g.* [3, 15, 16, 17], but they require a large number of $\boldsymbol{\Sigma}$ entries to be zero which may be too restrictive in some applications. In this work, we rather consider estimators in the second category based on shrinkage methods, introduced in [18]. In shrinkage methods, the considered estimators are convex combinations of an initial estimator and the identity matrix view as a regularization term. The construction of these estimators rely then on two main ingredients, the choice of the initial estimator to be regularized and the choice of the regularization parameter, or equivalently the weight of the identity matrix. As already mentioned, when aiming at robust inference, M-estimators are good initial basis. Following this line, the authors in [8] have proposed a shrinkage procedure, with an optimal shrinkage coefficient, for a particular case of M-estimators, called Tyler's estimator where the function $u(t)$ is set to p/t [19]. This choice of u is motivated by the fact that if $\boldsymbol{x}$ is elliptically distributed with mean $\boldsymbol{\mu}$, then the normalized vector $\boldsymbol{z} = (\boldsymbol{x} - \boldsymbol{\mu})/||\boldsymbol{x} - \boldsymbol{\mu}||$ follows an angular central Gaussian distribution. This approach has the advantage to be somewhat non-parametric and has shown a lot of merits in various settings [20, 21, 22]. Unfortunately, a serious limit is that it requires the mean $\boldsymbol{\mu}$ to be known in advance so that the shape of the distribution cannot be taken into account when estimating the mean. This point has been highlighted in [23], which proposes to estimate $\boldsymbol{\mu}$ assuming $\boldsymbol{x}$ follows a Cauchy distribution (*i.e.* a t-distribution with d.o.f. parameter equal to 1), and as a follow-up more recently in [24] with a generalization to any t-distributions. However, in contrast to [8], none of these papers provide an optimal shrinkage coefficient. Although the effect of tuning this coefficient may be important, the issue is usually eliminated either by searching in a finite grid of values [23, 24] or using cross-validation [25], in both cases at the cost of a higher computational complexity and time.

We aim at building on these previous approaches by providing both a flexible and optimal framework based on shrinked likelihood-based M-estimators. The

distribution of $\boldsymbol{x}$ is assumed to be elliptical so that the corresponding function u and the associated M-estimator can be derived straightforwardly from a maximum likelihood principle. We propose then a shrinkage version of this estimator with an explicit formula for the optimal shrinkage coefficient that depends on two moments of the radius of $\boldsymbol{x}$. Then, a complete inference procedure is proposed which does not require neither to pre-set the value of the mean nor that of the tail parameter. Explicit expressions of the optimal shrinkage coefficient are given for Gaussian and t-distributions and an algorithm for estimating both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is proposed. Experiments on simulated and real brain connectivity data illustrate the good behavior of the proposed method in comparison to other existing methods such as Tyler's estimator, graphical Lasso, etc.

The paper is organized as follows. Section 2 recalls definitions and the main properties of elliptical distributions and M-estimators. The optimal shrinkage problem is addressed in Section 3 with a general formula for the optimal shrinkage coefficient. In the following Section 4, the optimal parameter value is given in the case of multivariate t-distributions together with a practical algorithm to estimate both the mean and covariance matrix in a potentially low sample size setting. The proposed estimator and algorithm are illustrated on simulated and real data respectively in Section 5 and 6. A conclusion ends the paper. At last, all proofs and supplementary results are provided in Section 8 or as supplementary material.

2. Preliminaries

2.1. Elliptical distributions

A continuous random vector $\boldsymbol{x} \in \mathbb{R}^p$ follows a multivariate elliptical symmetric distribution if its probability density function (pdf) is of the form (see [26] or [27]):

$$p(\boldsymbol{x}) = C_{p,g} \, |\boldsymbol{\Sigma}|^{-1/2} g\left((\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \tag{1}$$

where $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ is the scale matrix with determinant $|\boldsymbol{\Sigma}|$, $\boldsymbol{\mu} \in \mathbb{R}^p$ is the location or mean vector, $C_{p,g}$ is a normalizing constant so that $p(\boldsymbol{x})$ integrates to one.

The non-negative function g is called the density generator and determines the shape of the pdf. Also, it is important to note that elliptical distributions have the stochastic representation $\boldsymbol{x} = \boldsymbol{\mu} + R\boldsymbol{\Lambda}\mathbf{U}$ [26], where R (called radius) is a non negative random variable, $\boldsymbol{\Lambda}$ is a $p \times p$ matrix so that $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top = \boldsymbol{\Sigma}$ and $\mathbf{U}$ is a p-dimensional random vector following a uniform distribution on the unit sphere in dimension p (R and $\mathbf{U}$ are independent). The radius R and the generator g are closely related. Indeed, according to Theorem 2.9 in [28], an elliptical distribution has a generator if and only if the random variable R has a density, and there exists a link between these two functions (Theorem 3 in [29] gives a similar result). Throughout this paper, we assume that our elliptical distribution has a generator, and the latter may be defined either by g, or by its radius R. This family encompasses a lot of well known particular cases, like the Gaussian distribution (with $g(t) = \exp(-t/2)$) and the Student distribution (also called t-distribution) with $\nu > 0$ degrees of freedom (with $g(t) = (1 + t/\nu)^{-(p+\nu)/2}$). Other examples include the Logistic [30], Kotz [31], Laplace [32] or Slash [33] distributions.

In this paper we consider the problem of the scale matrix estimation $\boldsymbol{\Sigma}$ from a set $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ of n independent and identically distributed (i.i.d.) p-dimensional observations drawn from an elliptical distribution (1). It is an important task both in the case of known or unknown location or mean vector $\boldsymbol{\mu}$. A lot of methods have already been proposed. For instance, [34] focused on the widely used sample covariance matrix $\widehat{\mathbf{S}} = n^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top$ as an estimator of $\boldsymbol{\Sigma}$ (the mean vector $\boldsymbol{\mu}$ is here considered as known, *i.e.* the data has previously been centered). However, being designed for the Gaussian distribution, this method is not suitable for the case of data with outliers. Moreover, it requires the existence of $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$, and this condition is not always fulfilled (see *e.g.* the Cauchy distribution). To overcome these difficulties, [19] proposed another estimator which is a particular case of Maronna's M-estimators [6] detailed in the next section.

To be more specific, we first provide below results for $\boldsymbol{\Sigma}$, but as done in many papers, *e.g.* [19, 8, 35, 36, 23], we propose in practice to estimate the

scale matrix normalized by its trace, also called the *shape* matrix. This both solves an intrinsic identifiability issue within the class of elliptical distributions [19] and helps the convergence of the estimation algorithm [8]. We provide, in Section 3.3, a detailed explanation that we can, without loss of generality, focus on the estimation of the shape matrix or trace-normalized scale matrix, with a trace set to p, *i.e.* $\mathbf{V} = p\boldsymbol{\Sigma}/\mathrm{tr}\boldsymbol{\Sigma}$. In particular, this is not the same as assuming directly that $\mathrm{tr}(\boldsymbol{\Sigma}) = p$. Note also that strictly speaking the covariance matrix when it exists is proportional to the scale matrix so that the term *covariance* is sometimes used abusively.

2.2. M-estimators and Tyler's estimator

Let $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be a set of n i.i.d. observations drawn from an elliptical distribution (1) with a known mean vector $\boldsymbol{\mu}$. D. E. Tyler [19] proposed a distribution-free estimator of the trace-normalized covariance matrix by working with the normalized observations $\boldsymbol{z}_i = \frac{\boldsymbol{x}_i - \boldsymbol{\mu}}{\|\boldsymbol{x}_i - \boldsymbol{\mu}\|^2}$. According to [37], each $\boldsymbol{z}_i$ follows the angular central Gaussian distribution:

$$p(\boldsymbol{z}) = \frac{\Gamma(p/2)}{2\pi^{p/2}} |\boldsymbol{\Sigma}|^{-1/2} \left(\boldsymbol{z}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{z}\right)^{-p/2}. \tag{2}$$

The maximum likelihood principle leads to an implicit estimator $\widetilde{\boldsymbol{\Sigma}}$, solution of

$$\widetilde{\boldsymbol{\Sigma}} = \frac{p}{n} \sum_{i=1}^{n} \frac{\boldsymbol{z}_i \boldsymbol{z}_i^\top}{\boldsymbol{z}_i^\top \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{z}_i}. \tag{3}$$

A fixed point algorithm is usually used to compute $\widetilde{\boldsymbol{\Sigma}}$ with a final normalization step to ensure $\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}) = p$. Tyler's estimator may then also be seen as a particular case of Maronna's M-estimator. Existence and uniqueness of $\widetilde{\boldsymbol{\Sigma}}$ are discussed in [19]. In particular, it is mentioned that the condition $n > p$ is required. Otherwise, according to [38], matrix $\widetilde{\boldsymbol{\Sigma}}$ is singular and this estimator is no longer suitable. In the case $p \geq n$, a regularized Tyler's estimator has been proposed, based on shrinkage methods [8] as specified in the following section.

2.3. Regularized Tyler's estimator

Inspired by the shrinkage method of Ledoit and Wolf [18], the authors in [8] extended Tyler's method to the high dimensional setting introducing the following regularized fixed point equations. The t^{th} iteration is indicated with index (t):

$$\widetilde{\mathbf{\Sigma}}^{(t+1)} = (1-\rho)\frac{p}{n}\sum_{i=1}^{n}\frac{\boldsymbol{z}_i\boldsymbol{z}_i^\top}{\boldsymbol{z}_i^\top\mathbf{\Sigma}^{(t)-1}\boldsymbol{z}_i} + \rho\mathbf{I}, \tag{4}$$

$$\mathbf{\Sigma}^{(t+1)} = \frac{p\,\widetilde{\mathbf{\Sigma}}^{(t+1)}}{\mathrm{tr}(\widetilde{\mathbf{\Sigma}}^{(t+1)})}. \tag{5}$$

Here $0 \leq \rho \leq 1$ is a constant which is called shrinkage coefficient. The case of $\rho = 0$ corresponds to the standard non regularized Tyler's estimator while $\rho = 1$ reduces the estimator to the identity matrix. The term $\rho\mathbf{I}$ ensures that the estimator is well-conditioned at each iteration. Both existence and uniqueness of the limit of the procedure (4)-(5) are proved in [8]. The choice of ρ is also discussed. As in [18], the authors in [8] proposed to find parameter ρ by minimizing the mean-squared error (MSE) between the true matrix $\mathbf{\Sigma}$ and the so-called "clairvoyant estimator":

$$\widetilde{\mathbf{\Sigma}}_\rho = (1-\rho)\frac{p}{n}\sum_{i=1}^{n}\frac{\boldsymbol{z}_i\boldsymbol{z}_i^\top}{\boldsymbol{z}_i^\top\mathbf{\Sigma}^{-1}\boldsymbol{z}_i} + \rho\mathbf{I}. \tag{6}$$

Thus, ρ is chosen as the solution ρ_T^* of:

$$\rho_T^* = \arg\min_{\rho}\ \mathbb{E}\left[\left\|(1-\rho)\frac{p}{n}\sum_{i=1}^{n}\frac{\boldsymbol{z}_i\boldsymbol{z}_i^\top}{\boldsymbol{z}_i^\top\mathbf{\Sigma}^{-1}\boldsymbol{z}_i} + \rho\mathbf{I} - \mathbf{\Sigma}\right\|_F^2\right], \tag{7}$$

where $||.||_F$ is the Frobenius norm. The solution can be seen as the value of ρ which minimizes the distance between the true $\mathbf{\Sigma}$ and its shrinked deformation. Following the above criteria, an explicit formula for ρ_T^* is obtained under the assumption $\mathrm{tr}(\mathbf{\Sigma}) = p$:

$$\rho_T^* = \frac{p^2 + (1-2/p)\mathrm{tr}(\mathbf{\Sigma}^2)}{(p^2 - np - 2n) + (n+1+2(n-1)/p)\mathrm{tr}(\mathbf{\Sigma}^2)}. \tag{8}$$

In the following developments, we propose to generalize this last result, to the case when $\boldsymbol{\mu}$ is not known and for all M-estimators when the data is sampled

from a specified elliptically symmetric distribution (1). Under a criterion similar to (7), we provide a closed-form expression for the optimal shrinkage coefficient.

3. Optimal shrinkage for M-estimators

Let $\delta_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}) = (\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ denote the Mahalanobis distance. In this section, we still suppose that $\boldsymbol{\mu}$ is known, and consider the class of Maronna's estimators [6, 35] satisfying :

$$\widetilde{\boldsymbol{\Sigma}} = m\left(\widetilde{\boldsymbol{\Sigma}}\right), \text{with} \tag{9}$$

$$m\left(\boldsymbol{\Sigma}\right) = \frac{1}{n}\sum_{i=1}^{n} u\left(\delta_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}_i)\right)\left(\boldsymbol{x}_i-\boldsymbol{\mu}\right)\left(\boldsymbol{x}_i-\boldsymbol{\mu}\right)^\top. \tag{10}$$

By taking $u(t) = p/t$ and $\boldsymbol{\mu} = \mathbf{0}$, we recover Tyler's estimator (3). Some other examples of functions u are $u(t) = 1$ [36], the Huber's function [39], or the Student maximum likelihood-based function $(p+\nu)/(t+\nu)$ [6]. As proposed in [36], in this paper we consider a regularized estimator:

$$\widetilde{\boldsymbol{\Sigma}}_{\alpha\beta} = \beta m\left(\boldsymbol{\Sigma}\right) + \alpha\mathbf{I}, \quad \alpha \geq 0, \beta \geq 0. \tag{11}$$

We define the following criteria, similar to (7), for the choice of α and β. The optimal (α^*, β^*) are chosen such as to minimize the MSE between the "clairvoyant estimator" $\widetilde{\boldsymbol{\Sigma}}_{\alpha\beta}$ and $\boldsymbol{\Sigma}$:

$$\mathbb{E}\left[\left\|\frac{\beta}{n}\sum_{i=1}^{n} u\left(\delta_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}_i)\right)\left(\boldsymbol{x}_i-\boldsymbol{\mu}\right)\left(\boldsymbol{x}_i-\boldsymbol{\mu}\right)^\top + \alpha\mathbf{I} - \boldsymbol{\Sigma}\right\|_F^2\right]. \tag{12}$$

Alternatively, another MSE criterion with $\beta = 1-\alpha$ has been considered in [36], for which it is also possible to give the optimal α value. This criterion is not further considered in our work but we provide the corresponding optimal α formula in Section 1 of the supplementary materials.

The optimal coefficients α^* and β^* minimizing (12) are given in Section 3.2, for elliptical distributions and for functions u derived from a maximum likelihood principle as explained in the following section.

3.1. Choice of function u

Natural choices for u are motivated by the maximum likelihood principle. Indeed, for an i.i.d. sample $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ from an elliptical distribution with stochastic representation $\boldsymbol{x} = \boldsymbol{\mu} + R\boldsymbol{\Lambda}\mathbf{U}$, the maximum likelihood estimator (MLE) of the scale matrix minimizes the negative log-likelihood function:

$$\mathcal{L}(\boldsymbol{\Sigma}) = -\frac{2}{n}\sum_{i=1}^{n} \ln\left(g\left(\delta_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}_i)\right)\right) - \ln|\boldsymbol{\Sigma}^{-1}|. \tag{13}$$

The previous equation leads to an implicit estimator of $\boldsymbol{\Sigma}$, obtained through a fixed point algorithm. However, this approach is no longer suitable if $p > n$. In that case, similarly to the approach of [36], the penalized cost function below can be considered:

$$\begin{aligned}\mathcal{L}_{\alpha\beta}(\boldsymbol{\Sigma}) = &-\beta\frac{2}{n}\sum_{i=1}^{n} \ln\left(g\left(\delta_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}_i)\right)\right) - \ln|\boldsymbol{\Sigma}^{-1}| \\ &+ \alpha\,\mathrm{tr}(\boldsymbol{\Sigma}^{-1}).\end{aligned} \tag{14}$$

In what follows, we suppose that the generator g, or equivalently the density of R, is differentiable. The solution $\widetilde{\boldsymbol{\Sigma}}_{\alpha\beta}$ which minimizes the penalized cost function $\mathcal{L}_{\alpha\beta}(\boldsymbol{\Sigma})$ can be expressed as:

$$\widetilde{\boldsymbol{\Sigma}}_{\alpha\beta} = \beta\frac{1}{n}\sum_{i=1}^{n} u(\delta_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}_i))(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top + \alpha\mathbf{I}, \tag{15}$$

with $u(t) = -2g'(t)/g(t)$.

We thus consider regularized M-estimators with function $u(t) = -2g'(t)/g(t)$. It is interesting to note that penalizations are linked to prior choice for $\boldsymbol{\Sigma}$ in a Bayesian framework. In (14) above, the $\mathrm{tr}(\boldsymbol{\Sigma}^{-1})$ penalization corresponds to an inverse Wishart prior where the scale matrix hyperparameter is the identity matrix. For a more general matrix hyperparameter $\mathbf{T}$, the penalty would be $\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{T})$ leading to a regularized estimator similar to (15) with a penalty term replaced by $\alpha\mathbf{T}$ and $\beta = 1 - \alpha$. Theorem 3.1 below can then be generalized to this case. This result and its proof are given in Section 2 of the supplementary materials.

3.2. Theoretical value of (α^*, β^*)

The proof of Theorem 3.1 below providing closed form formulas for the optimal shrinkage parameters (α^*, β^*) is given in Section 8.1. The sphericity measure ζ introduced in [40] is used to simplify the expressions. Sphericity $\zeta \in [0, p-1]$ is defined as

$$\zeta = \frac{p \operatorname{tr}(\mathbf{\Sigma}^2)}{\operatorname{tr}(\mathbf{\Sigma})^2} - 1 \,.$$

It measures how close $\mathbf{\Sigma}$ is to a scaled identity matrix, with $\zeta = 0$ when $\mathbf{\Sigma}$ is proportional to identity and $\zeta = p - 1$ when $\mathbf{\Sigma}$ has rank 1.

Theorem 3.1 (Optimal shrinkage coefficients)**.** *Let* $\boldsymbol{x} = \boldsymbol{\mu} + R\mathbf{\Lambda}\mathbf{U}$ *be a stochastic representation of the elliptically distributed* $\boldsymbol{x}$*, where* $\mathbf{\Lambda}\mathbf{\Lambda}^T = \mathbf{\Sigma}$ *and* R *is a positive random variable with a differentiable pdf and* $u(t) = -2g'(t)/g(t)$*. The oracle coefficients* (α^*, β^*) *which minimize* (12) *are*

$$\beta^* = \frac{n\zeta}{\left((\zeta+1)\left(n - 1 + \frac{2f_2}{p(p+2)}\right) + \frac{f_2}{p+2} - n\right)} \tag{16}$$

$$\alpha^* = (1-\beta^*)\frac{tr(\mathbf{\Sigma})}{p}, \tag{17}$$

where $f_2 = \mathbb{E}\left[u(R^2)^2 R^4\right]$ *and* $\beta^* \in [0,1]$*,* $\alpha^* \geq 0$ *for* $p \geq 2$*.*

These optimal values involve the quantity f_2. In [8], the particular choice $u(t) = p/t$ and assumption $\operatorname{tr}(\mathbf{\Sigma}) = p$ are made, hence $\alpha^* = 1 - \beta^*$ and $f_2 = p^2$. In the following proposition, other choices of u are considered and the corresponding f_2 values provided for several radius R. In the sequel, χ^2_k denotes the Chi-squared distribution with k degrees-of-freedom (d.o.f.) and $F_{p,\nu}$ denotes the Fisher distribution with d.o.f. parameters p and ν.

Proposition 3.1 (Some values of f_2)**.** *Let* $\boldsymbol{x} = \boldsymbol{\mu} + R\mathbf{\Lambda}\mathbf{U}$*, where* R *is a positive random variable with a differentiable pdf, and* $u(t) = -2g'(t)/g(t)$*.*

1. *If* R^2 *is distributed as* χ^2_p*, then* $\boldsymbol{x}$ *follows a Gaussian distribution,* $u(t) = 1$ *and therefore:*

$$f_2 \;=\; \mathbb{E}\left[u\left(R^2\right)^2 R^4\right] = p(p+2). \tag{18}$$

2. *If R^2 is distributed as $\left(\frac{1}{2\lambda}\chi^2_{\frac{2q+p-2}{s}}\right)^{1/s}$, then $\boldsymbol{x}$ follows a Kotz-type distribution, $u(t) = 2(1-q)/t + 2s\lambda t^{s-1}$ and therefore:*

$$f_2 \;=\; \mathbb{E}\left[u\left(R^2\right)^2 R^4\right] = p(p+2s) + 4s(q-1). \tag{19}$$

3. *If R^2 is distributed as $pF_{p,\nu}$, then $\boldsymbol{x}$ follows a t-distribution with $\nu > 0$ degrees-of-freedom, $u(t) = (p+\nu)/(t+\nu)$ and therefore :*

$$f_2 \;=\; \mathbb{E}\left[u\left(R^2\right)^2 R^4\right] = \frac{(\nu+p)(p+2)p}{p+\nu+2}. \tag{20}$$

The proof is provided in Section 8.2. The above formulas for f_2 are consistent with the fact that when ν goes to $+\infty$, the t-distribution tends to the Gaussian distribution and expression (20) tends to (18). A similar check can be done using that the Gaussian distribution is a particular case of Kotz-type distributions with $s = 1$, $q = 1$ and $\lambda = 1/2$.

Combining Theorem 3.1 and Proposition 3.1, the optimal shrinkage coefficients can be specified for the above distributions. In the following result, we restrict to the Gaussian and t-distributions.

Corollary 3.1 (Optimal shrinkage coefficients for multivariate Gaussian and t-distributions)**.** *The optimal shrinkage coefficients are given by,*

1. *For the Gaussian distribution:*

$$\beta^* = \frac{n\zeta}{(n+1)\zeta + p + 1} \tag{21}$$

$$\alpha^* = \frac{(\zeta+p+1)}{(n+1)\zeta+p+1}\;\frac{tr(\boldsymbol{\Sigma})}{p}. \tag{22}$$

2. *For the t-distribution with $\nu > 0$ degrees of freedom:*

$$\beta^* = \frac{n\zeta}{\left(n-1+2\frac{(\nu+p)}{(\nu+p+2)}\right)\zeta + \frac{(\nu+p)}{(\nu+p+2)}(p+2) - 1} \tag{23}$$

$$\alpha^* = \frac{\left(2\frac{(\nu+p)}{(\nu+p+2)} - 1\right)\zeta + \frac{(\nu+p)}{(\nu+p+2)}(p+2) - 1}{\left(n-1+2\frac{(\nu+p)}{(\nu+p+2)}\right)\zeta + \frac{(\nu+p)}{(\nu+p+2)}(p+2) - 1}\,\frac{tr(\boldsymbol{\Sigma})}{p}. \tag{24}$$

We now have explicit formulas for our optimal shrinkage coefficients in the t-distribution case including (for $\nu = 0$) the Tyler's coefficient ρ_T^* (equal to α^* when $\mathrm{tr}(\mathbf{\Sigma}) = p$) specified in (8) [8]. In practice, it still remains to compute estimations for $\mathrm{tr}(\mathbf{\Sigma})$, ζ and ν to get values for α^* and β^*. However, the need for $\mathrm{tr}(\mathbf{\Sigma})$ actually disappears with the use of a trace-normalized version of the scale matrix, namely $\mathbf{V} = p\mathbf{\Sigma}/\mathrm{tr}(\mathbf{\Sigma})$. This is detailed in the next section.

3.3. Trace-normalized scale matrices

Within the class of elliptical distributions, the scale matrix $\mathbf{\Sigma}$ suffers from some identifiability issue in the sense that the distribution defined in (1) is unchanged when $\mathbf{\Sigma}$ is replaced by $c\mathbf{\Sigma}$ and g by g_1 where $g_1(t) = c^{p/2}g(ct)$ for any fixed positive scalar c. In other words, any triplet $(\mu, \mathbf{\Sigma}, g)$ can be replaced by $(\mu, c\mathbf{\Sigma}, g_1)$. This change in g corresponds to changing R into $R_1 = R/\sqrt{c}$. The previous results apply for any new triplet $(\mu, c\mathbf{\Sigma}, g_1)$. It is easily checked that the sphericity ζ is unchanged and that for u_1 chosen following the maximum likelihood principle, *i.e.* $u_1(t) = -2\frac{g_1'(t)}{g_1(t)}$, then $u_1(t) = cu(ct)$ and the quantity f_2 is unchanged. Consequently, the optimal shrinkage coefficients in Theorem 3.1 are changed into $\alpha_1^* = c\alpha^*$ and $\beta_1^* = \beta^*$, which implies that using expression (15), $\widetilde{\mathbf{\Sigma}}_{\alpha_1^*\beta_1^*} = c\widetilde{\mathbf{\Sigma}}_{\alpha^*\beta^*}$. These two matrices have the same trace-normalization, *i.e.* $p\widetilde{\mathbf{\Sigma}}_{\alpha_1^*\beta_1^*}/\mathrm{tr}(\widetilde{\mathbf{\Sigma}}_{\alpha_1^*\beta_1^*}) = p\widetilde{\mathbf{\Sigma}}_{\alpha^*\beta^*}/\mathrm{tr}(\widetilde{\mathbf{\Sigma}}_{\alpha^*\beta^*})$. Applying this in particular with $c = p/\mathrm{tr}(\mathbf{\Sigma})$, it follows an algorithm detailed in Section 4 that provides an estimation of the trace-normalized $\mathbf{V} = p\mathbf{\Sigma}/\mathrm{tr}(\mathbf{\Sigma})$ for which we can use the fact that $\mathrm{tr}(\mathbf{V}) = p$ by construction. Considering $\mathbf{V}$, the expression for the optimal β is then unchanged and given by β^* in (16) while the expression for α^* simplifies into $\alpha^* = 1 - \beta^*$. Similarly, it is easy to check that the mean estimation, as proposed in Section 4, is not impacted by the trace-normalization.

4. Regularized trace-normalized scale matrix estimator for the multivariate *t*-distribution

In this section and in the sequel, we focus on the multivariate t-distribution case, and aim at estimating the mean vector $\boldsymbol{\mu}$ and $\mathbf{V} = p\mathbf{\Sigma}/\mathrm{tr}(\mathbf{\Sigma})$ using our shrinkage

methodology. The first step is thus to estimate the optimal shrinkage parameters which as explained above reduces to estimating β^* as given in (16) (with $\alpha^* = 1-\beta^*$). According to Corollary 3.1, β^* may be estimated using estimators of ζ (unchanged by the trace-normalization) and ν. The next paragraph gives some suitable estimators for these quantities.

4.1. Estimation of the optimal values α^* and β^*

To provide a numerical expression of α^* and β^*, the unknown quantities ζ and ν need to be estimated. For ζ, we use the estimator proposed in [40] defined as:

$$\hat{\zeta} = p\ \mathrm{tr}(\mathbf{S}^2) - \frac{p}{n} - 1, \text{ where} \tag{25}$$

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n} \frac{(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^\top}{\|\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}\|^2} \quad \text{and} \quad \widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i. \tag{26}$$

Note that in matrix $\mathbf{S}$ above, $\widehat{\boldsymbol{\mu}}$ is usually replaced by the median. Results were similar with the sample mean, which is then kept when comparing with other methods for fairness.

Regarding the d.o.f. parameter ν, for t-distributions, the norm $||\boldsymbol{x}||_2$ is regularly varying with tail index $1/\nu$ [41], *i.e.* :

$$\forall t > 0, \lim_{z\to+\infty} \frac{\mathbb{P}\left(||\boldsymbol{x}||_2 > tz\right)}{\mathbb{P}\left(||\boldsymbol{x}||_2 > z\right)} = t^{-1/\nu}. \tag{27}$$

Different estimators of the tail index are available in the literature, the most popular and widespread being the Hill estimator, introduced in [42]. By taking the inverse of the latter, we define an estimator $\hat{\nu}_{k_n}$ of ν :

$$\hat{\nu}_{k_n} = \left(\frac{1}{k_n}\sum_{i=1}^{k_n} \ln\left(\frac{||\boldsymbol{x}_{[i]}||_2}{||\boldsymbol{x}_{[k_n+1]}||_2}\right)\right)^{-1}, \tag{28}$$

where $\boldsymbol{x}_{[i]}$ denotes the ordered observations such that $||\boldsymbol{x}_{[1]}||_2 \geq \ldots \geq ||\boldsymbol{x}_{[k_n+1]}||_2 \geq \ldots \geq ||\boldsymbol{x}_{[n]}||_2$. The Hill estimator, and therefore $\hat{\nu}_{k_n}$, are related to a number k_n. On a theoretical point of view, the latter has to fulfill $k_n \to +\infty$ and $k_n/n \to 0$ as $n \to +\infty$, and leads to a compromise between a variable and biased estimation of ν. Indeed, a small k_n leads to a low biased and high variable estimation, while a large k_n increases the bias and reduces the variance (see Section 3.2

in [43] for details). Therefore, the choice of k_n, usually chosen as $\lfloor n^b \rfloor$ with $0 < b < 1$, is an important point, and is discussed in [44] in the t-distribution case. According to [44], a choice of $b \leq 4/(\nu + 4)$ is suitable. In the sequel, $b = 0.25$ is chosen, filling the last inequality for all $\nu \leq 12$, *i.e.* in most cases. Using the same setting as in Section 5 below, the quality of the proposed Hill-based estimator is illustrated in the supplementary materials Section 4, Figure 1. In this setting, the estimated value is overall always above the true value, with a reduced and small bias as n increases. One advantage of the proposed method, with respect to the more traditional method of moments, is not to require the existence of moments. Other estimation procedures are possible but were not investigated in this study.

4.2. Joint scale matrix and mean estimation

In the previous sections we assumed that the mean vector $\boldsymbol{\mu}$ was known. When the mean vector is unknown, the sample mean $\widehat{\boldsymbol{\mu}} = 1/n \sum_{i=1}^{n} \boldsymbol{x}_i$ can be used as an estimator but it is likely to perform poorly in the presence of outliers or simply in high-dimensional cases. The estimation of the scale matrix can then be severely degraded by an inaccurate estimation of the mean vector $\boldsymbol{\mu}$. To overcome this difficulty, we focus on the joint mean - scale matrix estimation as the solution of a system of equations of the following form defining Maronna's M-estimators [6]:

$$0 = \frac{1}{n} \sum_{i=1}^{n} h\left(\delta_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}_i)\right)(\boldsymbol{x}_i - \boldsymbol{\mu}), \tag{29}$$

$$\boldsymbol{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} u\left(\delta_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}_i)\right)(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top, \tag{30}$$

where the functions h and u satisfy a set of general assumptions stated in [6]. We propose to extend this approach to the high-dimensional/small sample size case by replacing the scale matrix by its regularized estimator. More specifically, the trace normalized scale matrix $\mathbf{V}$ is considered as explained in Section 3.3. It follows the iterative algorithm below for $\boldsymbol{\mu}$ and $\mathbf{V}$. Denoting the t^{th} iteration

with index (t):

$$\boldsymbol{\mu}^{(t+1)} = \frac{\sum_{i=1}^{n} h\left((\boldsymbol{x}_i - \boldsymbol{\mu}^{(t)})^\top \mathbf{V}^{(t)-1}(\boldsymbol{x}_i - \boldsymbol{\mu}^{(t)})\right)\boldsymbol{x}_i}{\sum_{i=1}^{n} h\left((\boldsymbol{x}_i - \boldsymbol{\mu}^{(t)})^\top \mathbf{V}^{(t)-1}(\boldsymbol{x}_i - \boldsymbol{\mu}^{(t)})\right)}, \tag{31}$$

$$\widetilde{\mathbf{V}}^{(t+1)} = \frac{\beta}{n}\sum_{i=1}^{n} u\Big((\boldsymbol{x}_i - \boldsymbol{\mu}^{(t+1)})^\top \mathbf{V}^{(t)-1}(\boldsymbol{x}_i - \boldsymbol{\mu}^{(t+1)})\Big) \times (\boldsymbol{x}_i - \boldsymbol{\mu}^{(t+1)})(\boldsymbol{x}_i - \boldsymbol{\mu}^{(t+1)})^\top + (1-\beta)\mathbf{I}, \tag{32}$$

$$\mathbf{V}^{(t+1)} = p\frac{\widetilde{\mathbf{V}}^{(t+1)}}{\mathrm{tr}(\widetilde{\mathbf{V}}^{(t+1)})}. \tag{33}$$

The optimal β coefficient, for any distribution in the elliptical family, can be computed using expression (16). This value involves ζ and the parameters defining the distribution of the radius R. While ζ can be estimated using expression (25), the estimation of the R distribution parameters may not be obvious and requires additional examination. One possible solution is to set these parameters to fixed values representing the prior knowledge about the distribution of R. Another solution is to choose the parameters values corresponding to the heaviest tail case within the chosen distribution subclass. For example, for the multivariate t-distribution, the authors in [23] propose to set $\nu = 1$, in other words, to focus on the Cauchy distribution, which corresponds to a heavy-tail representative among the subclass of t-distributions. To keep more flexibility, these parameters could also be estimated using a maximum likelihood approach, possibly using the Expectation-Maximization (EM) algorithm. The EM algorithm is tractable for a subclass of the elliptical family referred to as Gaussian scale mixtures (GSM) [45]. GSM distributions include the generalized Gaussian, the multivariate t-distribution, the Pearson type VII distribution, etc.

Hereafter, Algorithm 1 provides a simple algorithm for the joint mean - scale matrix estimation for the multivariate t-distribution. This algorithm does not require an additional step for the estimation of the d.o.f. ν in contrast to an EM algorithm implementation that would iteratively update ν as well as $\boldsymbol{\mu}$ and $\mathbf{V}$. Instead, in Algorithm 1, ν is evaluated once through the Hill estimator (28) and used then for both the estimation of α^*, β^* and that of $\boldsymbol{\mu}$ and $\mathbf{V}$.

We do not provide a convergence proof for Algorithm 1 but we note that

Algorithm 1 Small sample size trace normalized scale matrix and mean estimation for a multivariate t-distribution

1: Initialize $\mathbf{V}$ to $\mathbf{V}^{(0)}$ an arbitrary positive definite matrix and $\boldsymbol{\mu}$ to $\boldsymbol{\mu}^{(0)}$ an arbitrary vector

2: Estimate d.o.f. $\widehat{\nu}$ using (28) and $\widehat{\beta}^*$ using (23) and (25), and $\widehat{\alpha}^* = 1 - \widehat{\beta}^*$.

3: Iterate the following steps until convergence:

3.1: Update the mean vector $\boldsymbol{\mu}$ as :

$$\boldsymbol{\mu}^{(t+1)} = \frac{\sum_{i=1}^{n} \bar{w}_i^{(t+1)} \boldsymbol{x}_i}{\sum_{i=1}^{n} \bar{w}_i^{(t+1)}},$$
$$\text{where } \bar{w}_i^{(t+1)} = \frac{\widehat{\nu} + p}{\widehat{\nu} + (\boldsymbol{x}_i - \boldsymbol{\mu}^{(t)})^\top \left(\mathbf{V}^{(t)}\right)^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}^{(t)})}. \tag{34}$$

3.2: Compute matrix $\widetilde{\mathbf{V}}^{(t+1)}$ as :

$$\widetilde{\mathbf{V}}^{(t+1)} = \widehat{\beta}^* \frac{p + \widehat{\nu}}{n}$$
$$\times \sum_{i=1}^{n} \frac{(\boldsymbol{x}_i - \boldsymbol{\mu}^{(t+1)})(\boldsymbol{x}_i - \boldsymbol{\mu}^{(t+1)})^\top}{(\boldsymbol{x}_i - \boldsymbol{\mu}^{(t+1)})^\top \left(\mathbf{V}^{(t)}\right)^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}^{(t+1)}) + \widehat{\nu}} + \widehat{\alpha}^* \mathbf{I}. \tag{35}$$

3.3: Update $\mathbf{V}$ as the trace-normalized $\widetilde{\mathbf{V}}^{(t+1)}$:

$$\mathbf{V}^{(t+1)} = p \, \frac{\widetilde{\mathbf{V}}^{(t+1)}}{\mathrm{tr}(\widetilde{\mathbf{V}}^{(t+1)})}.$$

step 3.3 in Algorithm 1 is crucial for convergence. Indeed, considering only the part involving $\mathbf{V}$ assuming $\boldsymbol{\mu}$ fixed, the convergence of the iterations derives from concave Perron-Frobenius theory following the proof given in [8] (Appendix VIII.A) with only minor adaptations. It appears in this proof that the normalization in step 3.3 is essential. In practice, we did observe unstable behaviors when this normalization was not imposed.

5. Results on simulated data

In this section we conduct a simulation study to illustrate the performance of the proposed shrinkage approach through Algorithm 1. In our experiments, an autoregressive (AR) covariance structure is considered :

$$(\boldsymbol{\Sigma})_{ij} = r^{|i-j|}, \ r \in (0,1). \tag{36}$$

When r tends to 0, then $\boldsymbol{\Sigma}$ is close to an identity matrix; when r tends to 1, then $\boldsymbol{\Sigma}$ is close to a singular matrix of rank 1. As pointed out earlier, the choice of b for computing the degrees of freedom parameter $\hat{\nu}_{k_n}$ in (28) is an important issue. The optimal theoretical value of this parameter depends on the true ν through formula $b \approx 4/(\nu+4)$. Here we choose $b = 0.25$ which corresponds to a suitable value for $\nu \leq 12$. This allows a robust estimation of ν both for heavy-tailed distributions (*e.g.* with $\nu = 1$ corresponding to a Cauchy distribution) and light-tailed distributions (*e.g.* $\nu = 12$).

In the first experiment, we simulate data from a multivariate t-distribution in dimension $p = 50$, with the following different d.o.f. parameter $\nu \in \{1, 2, 3, 6, 10\}$ and various AR schemes with $r \in \{0.1, 0.7, 0.9\}$. The mean $\boldsymbol{\mu}$ is set to the vector with all components equal to 5. For each pair of parameters (ν, r) and for a sample size n varying from 5 to 50, 100 data sets are generated leading to estimations $\widehat{\boldsymbol{\mu}}_s$, $\widehat{\boldsymbol{\Sigma}}_s$ for $s = 1$ to 100. The performance of a method is then assessed using the normalized mean square-error (NMSE) for both $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$. Since $\mathrm{tr}(\boldsymbol{\Sigma}) = p$, we can directly compare the true $\boldsymbol{\Sigma}$ with our estimations. In more

general cases, we would compare to the true trace-normalized $\frac{p\boldsymbol{\Sigma}}{\mathrm{tr}(\boldsymbol{\Sigma})}$ instead:

$$NMSE(\boldsymbol{\Sigma}) = \frac{\mathbb{E}\left\{\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_F^2\right\}}{\|\boldsymbol{\Sigma}\|_F^2} \approx \frac{1}{100}\sum_{s=1}^{100} \frac{\|\widehat{\boldsymbol{\Sigma}}_s - \boldsymbol{\Sigma}\|_F^2}{\|\boldsymbol{\Sigma}\|_F^2}, \tag{37}$$

$$NMSE(\boldsymbol{\mu}) = \frac{\mathbb{E}\left\{\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_F^2\right\}}{\|\boldsymbol{\mu}\|_F^2} \approx \frac{1}{100}\sum_{s=1}^{100} \frac{\|\widehat{\boldsymbol{\mu}}_s - \boldsymbol{\mu}\|_F^2}{\|\boldsymbol{\mu}\|_F^2}. \tag{38}$$

For each data set, five different algorithms are then used leading to five different estimators of $\boldsymbol{\Sigma}$ and four different estimators of $\boldsymbol{\mu}$. In this setting, since $\mathrm{tr}(\boldsymbol{\Sigma}) = p$, estimating ζ is equivalent to estimating $\mathrm{tr}(\boldsymbol{\Sigma}^2)$, for which propose to use $p^2\mathrm{tr}(\mathbf{S}^2) - \frac{p^2}{n}$, where $\mathbf{S}$ is defined in (26). The derivation of this expression is explained in more details in Section 3 of the supplementary materials.

Then, the following methods were used for the $NMSE(\boldsymbol{\Sigma})$ comparison:

- The algorithm proposed in [8] that estimates $\boldsymbol{\Sigma}$ on the data centered by the sample mean (sm), referred to as "Tyler sm". As a plug-in estimate for $\mathrm{tr}(\boldsymbol{\Sigma}^2)$ to compute ρ_T^* in (8) the authors use $p^2\mathrm{tr}(\mathbf{S}^2)$.

- For a fair comparison with "Tyler sm", we run only steps 3.2 and 3.3 of Algorithm 1, setting $\boldsymbol{\mu}$ to the sample mean and estimating $\boldsymbol{\Sigma}$ on the data centered by the sample mean. Moreover, for the estimation of β^* in (23) we also use $p^2\mathrm{tr}(\mathbf{S}^2)$ for $\mathrm{tr}(\boldsymbol{\Sigma}^2)$ instead of $p^2\mathrm{tr}(\mathbf{S}^2) - p^2/n$. This algorithm is denoted as "t-dist sm".

- To demonstrate the impact of estimating $\boldsymbol{\mu}$ iteratively instead of using the sample mean, we run Algorithm 1 where $\mathrm{tr}(\boldsymbol{\Sigma}^2)$ in $\widehat{\beta}^*$ is estimated using $p^2\mathrm{tr}(\mathbf{S}^2)$. This algorithm is referred to as "t-dist".

- We run the full Algorithm 1 referred to as "t-dist-2". Both $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ are computed iteratively and β^* is set with ζ in (25).

- At last we run Algorithm 1 with the theoretical oracle $\beta^*(\boldsymbol{\Sigma}, \nu)$ in (23) obtained with the true values of $\boldsymbol{\Sigma}$ and ν denoted as "Oracle".

To illustrate the impact of the estimation of $\boldsymbol{\mu}$, a comparison of the $NMSE(\boldsymbol{\mu})$ values is also provided for the four estimators resulting from the above algorithms, namely :

- the sample mean denoted by "sm";
- the estimation of $\boldsymbol{\mu}$ resulting from the "t-dist" algorithm;
- the estimation of $\boldsymbol{\mu}$ resulting from the "t-dist-2" algorithm (Algorithm 1);
- the estimation of $\boldsymbol{\mu}$ from the "Oracle" algorithm.

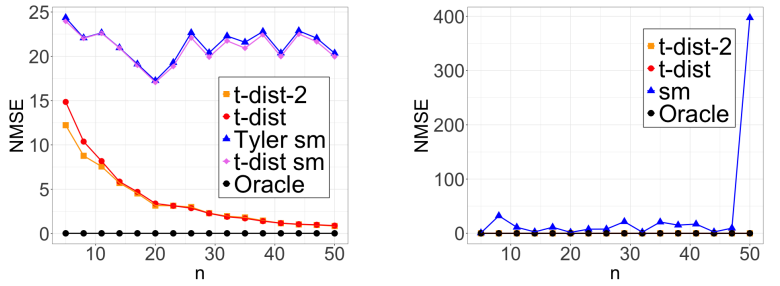
The results are illustrated in Figure 1 while more complete results are shown in Figures 2 and 3 in the supplementary materials, Section 4. In all scenarios, the proposed "t-dist" and "t-dist-2" algorithms are consistently the closest to the "Oracle" procedure, which as expected always provides the best results with the minimal $NMSE(\boldsymbol{\Sigma})$ and $NMSE(\boldsymbol{\mu})$. More specifically, the "t-dist" and "t-dist-2" performances are very close to the ideal oracle estimator with increasing n. However, it can be noted that "t-dist-2" outperforms "t-dist" significantly, especially for $NMSE(\boldsymbol{\Sigma})$ when the sample sizes are small. Since "t-dist" and "t-dist-2" differ only in the way they estimate $\mathrm{tr}(\boldsymbol{\Sigma}^2)$ for β^* in (23), the performance loss of "t-dist" over "t-dist-2" can be attributed to the bias of the estimator $p^2\mathrm{tr}(\mathbf{S}^2)$ as compared to $p^2\mathrm{tr}(\mathbf{S}^2) - p^2/n$.

To illustrate the importance of estimating $\boldsymbol{\mu}$ and ν, we can focus on the comparison of the "t-dist", "t-dist sm" and "Tyler sm" algorithms, which all use the same estimator $p^2\mathrm{tr}(\mathbf{S}^2)$ for $\mathrm{tr}(\boldsymbol{\Sigma}^2)$. For both "Oracle" and "t-dist" methods, $NMSE(\boldsymbol{\Sigma})$ steadily decreases as n increases. In contrast, the "Tyler sm" and "t-dist sm" algorithms show in some cases irregular NMSE curves due to a bad estimation of $\boldsymbol{\mu}$ by the sample mean. As illustrated in Figure 1(b) and supplementary Figure 3(a,b,c), this is particularly so for small d.o.f. like $\nu = 1$. This confirms the potential limits of methods that do not estimate $\boldsymbol{\mu}$ accurately as pointed out in [23]. As regards the impact of ν, it is not easy to illustrate separately the effect of ν from that of $\boldsymbol{\mu}$. To do so we consider the comparison of "Tyler sm" and "t-dist sm" algorithms (Figures 2 and 3). In the first step, both procedures employ the sample mean vector to center the original data. In the second step, the covariance matrix is computed using the iteration given in (4) for Tyler's estimator with $\widehat{\rho}_{T^*}$ defined in (8) and iteration given in step 3.2

for "t-dist sm" with $\widehat{\beta}^*$ in (23). Tyler's procedure can be viewed as an extreme case of the multivariate t-distribution with $\nu = 0$. Accordingly, we expect that the difference between the two estimators becomes more significant for larger values of ν. The better results provided by "t-dist sm" over "Tyler sm" are shown more specifically on Figure 3 for $\nu = 10$ where the main gains appear for small sample sizes (Figure 3(b)). Algorithm "t-dist sm" also outperforms "Tyler sm" for smaller d.o.f. like $\nu = 1$ but with an increasing gain as n becomes larger (see Figure 2(c,d)). However, for small ν the differences are somewhat less visible due to larger NMSEs coming mainly from a bad estimation of $\boldsymbol{\mu}$. When ν is large, the resulting curve for "t-dist sm" coincides with the one for "t-dist". This is not surprising as the t-distribution tends to the Gaussian distribution when ν tends to ∞ so that the mean vector in (34) becomes closer to the sample mean. This result is confirmed by Figure 1(d,f), supplementary Figure 3(j,k,l) for $\nu = 6$ and supplementary Figure 3(m,n,o) for $\nu = 10$. Overall estimating the d.o.f. parameter ν in step 3.2 is important, especially in the small sample size regime, and more generally because it allows a better estimation of the mean vector $\boldsymbol{\mu}$ which seems to have a critical impact of the covariance structure estimation. For both aspects, our proposed procedure improves over the regularized Tyler's algorithm. Regarding the impact of the choice of r, it does not seem to lead to significantly different conclusions (see Figures 2 and 3 in supplementary materials).

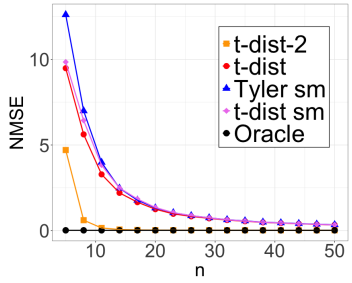
6. Application to brain connectivity data.

Robust estimation of covariance matrices is especially needed for real data where we know that Gaussian hypotheses are generally not true. This is the case for the inference of brain connectivity. Thanks to non invasive neuroimaging, brain recordings are now available to follow the activity of the brain during a task or at rest. Using functional magnetic resonance imaging (fMRI), one volume of the brain is acquired every one second or less for several minutes. Usually each volume is composed of thousands of voxels that are gathered into a set of

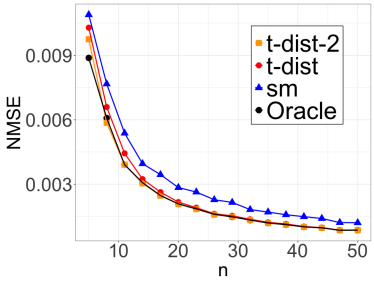


(a) $NMSE(\boldsymbol{\Sigma})$, $\nu = 1$

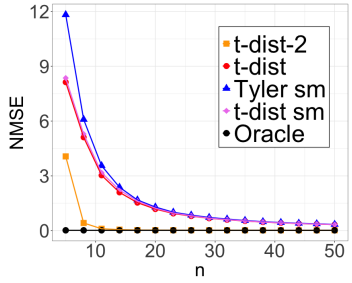
(b) $NMSE(\boldsymbol{\mu})$, $\nu = 1$



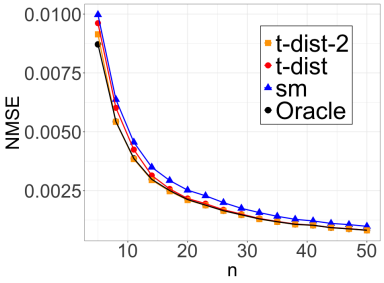
(c) $NMSE(\boldsymbol{\Sigma})$, $\nu = 6$



(d) $NMSE(\boldsymbol{\mu})$, $\nu = 6$

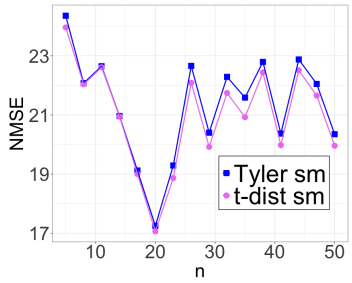


(e) $NMSE(\boldsymbol{\Sigma})$, $\nu = 10$

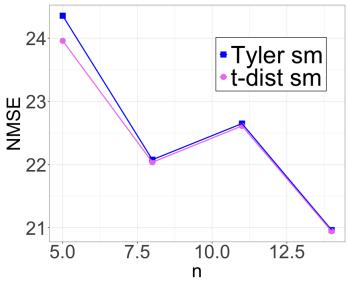


(f) $NMSE(\boldsymbol{\mu})$, $\nu = 10$

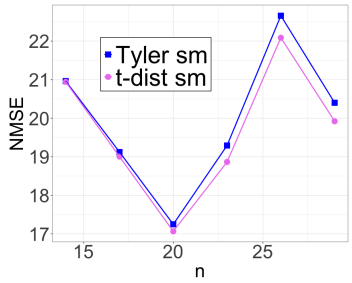
Figure 1: Multivariate t-distribution with AR(r) covariance structure ($p = 50$, $r = 0.1$, $\nu \in \{1, 6, 10\}$ and $\boldsymbol{\mu}$ is set to a vector of 5). Normalized mean squared-errors for $\boldsymbol{\Sigma}$ (first column) and $\boldsymbol{\mu}$ (second column) are computed over 100 simulated samples of n observations each with n varying from 5 to 50.



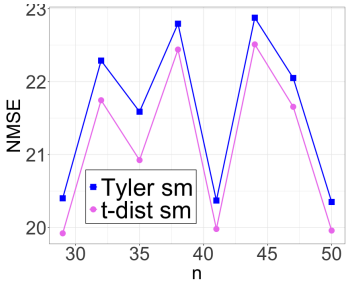
(a) $r = 0.1,\ \nu = 1,\ 5 \leq n \leq 50$



(b) $r = 0.1,\ \nu = 1,\ 5 \leq n \leq 15$

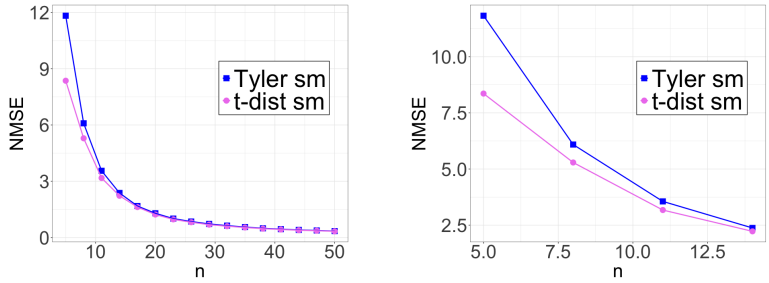


(c) $r = 0.1,\ \nu = 1,\ 15 \leq n \leq 30$



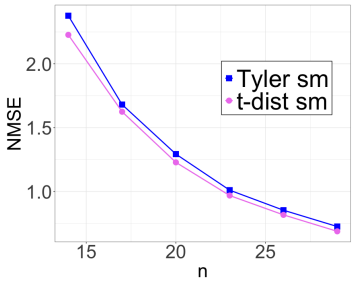
(d) $r = 0.1,\ \nu = 1,\ 30 \leq n \leq 50$

Figure 2: Multivariate t-distribution with AR(r) covariance structure ($p = 50$, $r = 0.1$, $\nu = 1$ and $\boldsymbol{\mu}$ is set to a vector of 5): comparison of $NMSE(\boldsymbol{\Sigma})$ for "Tyler sm" and "t-dist sm" algorithms when the mean is fixed to the sample mean. Normalized mean squared-errors are computed over 100 simulated samples of n observations each. Varying values of n from 5 to 50 (a) are also plotted separately for better layout in (b,c,d).

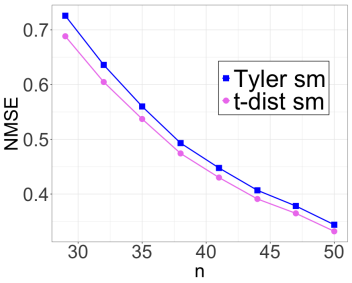


(a) $r = 0.1, \nu = 10,\ 5 \leq n \leq 50$

(b) $r = 0.1,\ \nu = 10,\ 5 \leq n \leq 15$



(c) $r = 0.1, \nu = 10,\ 15 \leq n \leq 30$



(d) $r = 0.1, \nu = 10,\ 30 \leq n \leq 50$

Figure 3: Moderate-tailed t-distribution ($\nu = 10$) with AR(r) covariance structure ($p = 50$, $r = 0.1$ and $\boldsymbol{\mu}$ is set to a vector of 5): comparison of $NMSE(\boldsymbol{\Sigma})$ for "Tyler sm" and "t-dist sm" algorithms when the mean is fixed to the sample mean. Normalized mean squared-errors are computed over 100 simulated samples of n observations each. Varying values of n from 5 to 50 (a) are also plotted separately in (b,c,d).

hundreds of parcels or brain regions. Each brain region is then associated to a time series. These data are still too complex to provide an easy visualisation and interpretation. Brain connectivity graphs or networks are then constructed by defining nodes as brain regions and edges as connections between time series associated to these brain regions. This allows a spatio-temporal modeling of the brain while functioning.

In this application, to quantify the links between time series, edges are associated to partial correlations read on inverse covariance matrices. Inference of brain connectivity graphs depends then on accurate estimation of covariance or precision matrices. To compare objectively different methods, we use a test-retest dataset selected from a larger dataset publicly released as part of the Human Connectome Project (HCP), WU-Minn Consortium (`https://www.humanconnectome.org/`).
More details can be found in [46]. We select 100 subjects who have been scanned twice in two different sessions of about 15 minutes each. These two sessions are divided into two sub-sessions of half duration and denoted respectively by S_{11}, S_{12} for session 1 and S_{21}, S_{22} for session 2. Shorter sessions are more common in practice and they represent a higher challenge for the tested approaches. Following [46], fMRI time series are analyzed through their wavelets decompositions providing vectors of wavelets coefficients resulting in datasets of size $n = 547$. However, the wavelets coefficients being not independent, the actual effective sample size is evaluated as being only 37. The number of brain regions is set to $p = 90$ based on a commonly used parcellation of the brain into 90 regions [47].

Before processing, several statistical tests were performed to check the heavy-tailed non-Gaussian nature of the time series. Multivariate Gaussianity tests exist but they are generally designed for conventional low-dimensional data. Proposals have been made in the small sample size setting, *e.g.* [48], based on robust estimations of the mean and covariance matrix but they could not be used here as this is precisely the goal of the paper to provide such estimations. However, since marginals of Gaussian vectors are all Gaussian, Shapiro-Wilk

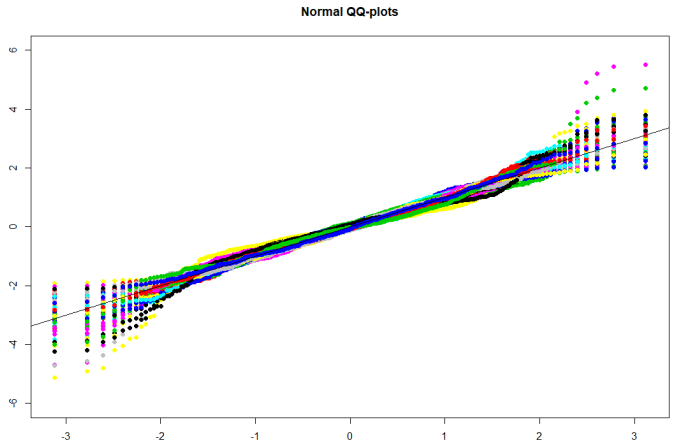


Figure 4: Quantile quantile plots for the time series corresponding to region 1 (first marginal), in sub-session S_{11}. Each color represents one of the 100 subjects. Empirical quantiles are plotted against theoretical Gaussian quantiles.

tests were performed on 1D marginals. They resulted for most subjects in the rejection of the Gaussian hypothesis. As an illustration, the histograms of the p-values for the first and 90^{th} marginals (*i.e.* brain regions 1 and 90), in the first sub-session S_{11} for the 100 subjects, are shown in supplementary Figure 4. For 50% of the subjects, normality is rejected at a significance level lower than 0.005 for the first region and lower than 0.007 for the 90^{th} region. In both cases, the average p-value was of 0.05. For an additional visual illustration, quantile-quantile plots, for each of the 100 subjects, session S_{11}, region 1 (first marginal), are shown in Figure 4 below. A similar plot for the 90^{th} region is shown in supplementary Figure 5. The departure from the straight line indicates non Gaussianity and the presence of an heavy tail. The adequation to an elliptical distribution was then not formally tested but 2D scatter plots provided visual hints showing acceptable elliptical shapes. The first and 90^{th} marginals of the first 16 subjects are provided in supplementary Figure 6. Similar results and plots were obtained for most of the subjects and sessions.

The reliability of each structure learning approach is then evaluated by mea-

suring for each subject, the graph properties differences for the four pairs of sub-sessions coming from different sessions, (S_{11}, S_{21}), (S_{11}, S_{22}), (S_{12}, S_{21}) and (S_{12}, S_{22}). For each dataset, five different approaches are considered:

- Sample partial correlation denoted as "sample pc".
- Shrinkage using the Ledoit-Wolf's estimator in [18] referred as "lw".
- Maximum-likelihood for a t-distribution using EM [49] defined as "t-dist em".
- Graphical Lasso using [11] which we refer to as "glasso" (the tuning parameter λ in "glasso" algorithm for each subject in each sub-session was obtained by cross-validation with $k = 3$ folds).
- Our shrinkage Algorithm 1 for a t-distribution. We call this method "t-dist-2 shrink".

In contrast to simulated data experiments, results with Tyler's estimators are not shown here because they provide results similar to the Ledoit-Wolf's estimator.

In order to produce graphs with a fixed comparable number of edges, we apply soft-thresholding to each obtained matrix. For each subject in each sub-session, we then obtain an adjacency matrix that defines an unweighted graph for which a graph metric called *global efficiency* is computed. This metric is related to the communication efficiency of a node i with all other nodes (detailed information can be found in [50]). If $G = (V, E)$ denotes a graph with V as its set of p vertices and E as its edge set, the global efficiency $Eglob_i$ is defined as the inverse of the harmonic mean of the set of the minimum path lengths L_{ij} between node $i \in V$ and all other nodes $j \in V$ in the graph:

$$Eglob_i = \frac{1}{p-1} \sum_{j \in V} \frac{1}{L_{ij}}. \tag{39}$$

Here p is the number of brain regions. Then by averaging these global efficiency values over all nodes, one value of this parameter is derived for a given graph.

Consequently, for a given pair of sub-sessions, 200 global efficiency values, one per each subject in each sub-session, are computed for a given pre-set percentage of edges in the graphs.

6.1. Brain connectivity graphs

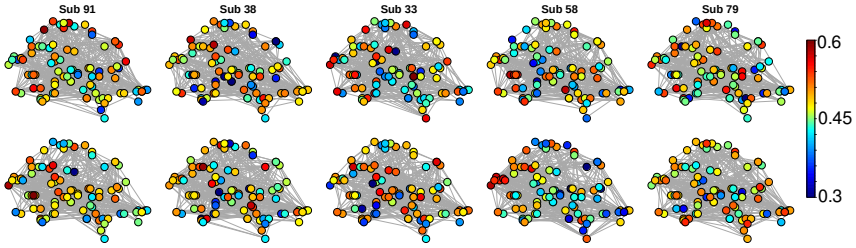
As an illustration, Figure 5 displays, for five subjects and sub-sessions S_{11} and S_{22}, the global efficiency computed for each region of the brain with either sample partial correlations or partial correlations using our shrinkage method. The global efficiency values are on average between 0.35 for the right Precentral region and 0.53 for the left Putamen region. Regions with high global efficiency greater than 0.5 include the post Cingulum, Amygdala, Frontal Middle Orbital, Occipital Inferior and Thalamus regions. Whereas regions with low efficiency less than 0.4 include the Precentral, Postcentral, Parietal Superior and Frontal Superior regions. The qualitative comparison between the two sub-sessions highlights a higher similarity and reproducibility between sessions with our shrinkage method "t-dist-2 shrink", Figure 5 (b), than with sample partial correlation "sample pc", Figure 5 (a). Similar results are observed for other subjects and other pairs of sub-sessions.

6.2. Test-retest reliability

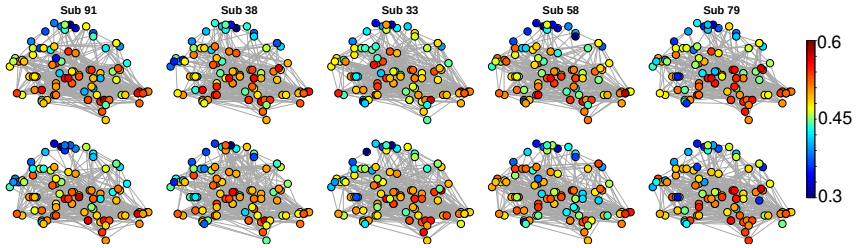
To quantify more specifically the differences between the five tested methods, we evaluate their ability to provide similar results between two sessions via the so-called intraclass correlation coefficient (ICC) between the sessions. Using the global efficiency values for each subject in each session, we compute their within-subject (s_w) and between-subject (s_b) mean square differences, as detailed in the Appendix of [46]. In our case, with two sessions, the ICC is then given by

$$\mathrm{ICC} = \frac{s_b - s_w}{s_b + s_w}. \tag{40}$$

When for each subject, similar global efficiency values are found in the two sessions, then the ICC is close to 1 and the reliability is high. In contrast, ICC is close to 0 when the reliability is low. The ICC may take negative values when



(a) sample pc



(b) t-dist-2 shrink

Figure 5: Global efficiency per brain region (node) for brain graphs with 10% of edges. Five subjects are displayed (columns). Two sessions S_{11} (first and third rows) and S_{22} (second and fourth rows) are compared using (a) sample partial correlations and (b) our proposed shrinkage approach. A high global efficiency means that the node/region is well connected to other nodes. Hubs generally show a high global efficiency. In contrast, a low global efficiency means that the shortest path length is large, and is typical of less connected nodes.

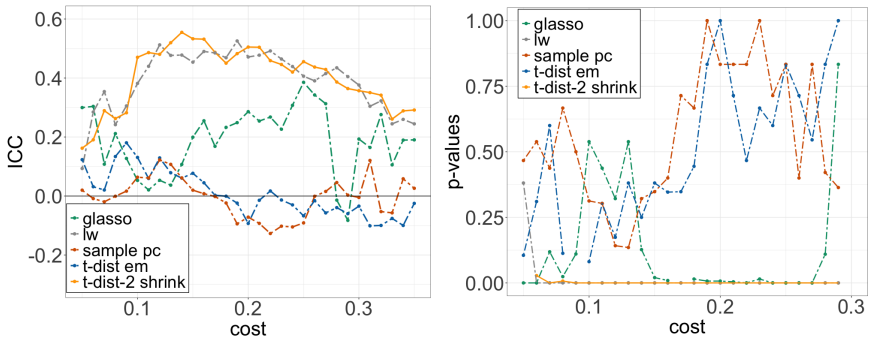
the variance within subjects is larger than between subjects. This is due to statistical errors given a particular dataset and should be considered as a non reliable estimation.

Figure 6 represents, for sessions S_{11} and S_{22}, ICC values with respect to the pre-set percentage of edges in the graphs, referred to as the cost. Shrinkage methods ("lw" and "t-dist-2 shrink") have the largest range of costs where ICC is above 0.4, Figure 6(a). This is confirmed by the computation of p-values, Figure 6 (b), where p-values allow to check whether the ICC is significantly larger than zero. The ICC computed using empirical partial correlation and EM for a t-distribution are nearly equal to 0 for any cost showing a poor reliability of global efficiency. Glasso is also showing poor performance because of the difficulties to choose the regularisation parameters. For additional confirmation that shrinkage schemes provide better results than sample partial correlation, we further check s_b values as an increase in s_b may artificially increase ICC values. Figure 7 displays s_b and s_w values obtained with the two shrinkage methods against the sample partial correlation values. Figure 7 (b) shows clearly that s_b behaves similarly in all three methods which allows then a fair comparison between them. In contrast, Figure 7 (a) shows that there is a clear decrease of s_w using methods based on shrinkage. This confirms that shrinkage estimators such as "lw" and "t-dist-2 shrink" show very good similar performances and improve the reliability of global efficiency on this test-retest dataset.

Similar conclusions hold for the other pairs of sessions. The corresponding ICC and p-values plots can be found in Section 5 of the supplementary materials, Figures 7 and 8.

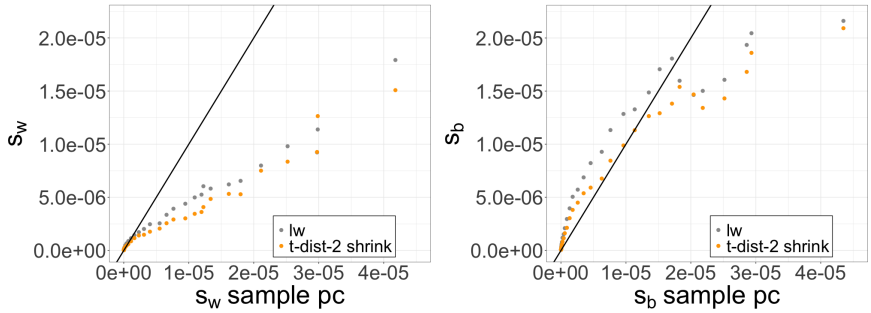
7. Conclusion

In this paper, we address the issue of robust covariance matrix estimation in settings where the sample size is small compared to the number of parameters and the mean is not known a priori. Elliptical distributions are considered to improve robustness. In particular, we focus on Student's t-distributions for their



(a) ICC (b) p-value

Figure 6: Intraclass correlation coefficient (ICC) between two fMRI sessions (S_{11} and S_{22}) (a) and associated p-values (b) with respect to the pre-set percentage of edges in the graphs (cost). The ICC values are shown for the various estimators considered in this study. The larger the ICC, the higher the consistency between the two sessions and the higher the estimator reliability.



(a) Within-subject variances (s_w) (b) Between-subject variances (s_b)

Figure 7: Within-subject s_w (a) and between-subject s_b (b) variances, for sessions S_{11} and S_{22}, using different estimators. Values of s_w and s_b for two shrinkage methods ("lw" and "t-dist-2 shrink") are plotted against values obtained using sample partial correlation. The black line indicates the line of equal values. Red and grey dots correspond to varying percentages of edges in the graphs.

ability to model heavy tails and to handle outliers. A regularisation approach based on shrinkage is then used to face the relative lack of data. These two aspects are combined and lead to a penalized maximum likelihood based estimator assuming the observations follow a multivariate Student's t-distribution. The proposed approach is showed to fulfill theoretical results for more general elliptical distributions, and it has the advantage to be implemented easily in practice.

Among regularized robust estimators, the proposed estimator has several desirable properties: 1) the penalization level or regularizing coefficient is not tuned manually but estimated via a closed-form formula deriving from a minimum mean squared-error principle, 2) prior knowledge on the mean and degree-of-freedom parameter values are not needed and both these parameters can be estimated in a data driven way, at last 3) the efficient algorithm that is derived shows good estimation accuracy when compared to the standard Tyler's estimator on simulated data and to additional standard methods on real data. In particular our experiments confirm the importance of a good estimation of the mean and the potential advantage of methods that aim at estimating both the mean and covariance matrix.

8. Proof of the main results

8.1. Proof of Theorem 3.1

Using the same notation as in [8], we define matrix $\widetilde{\mathbf{C}}$ as:

$$\widetilde{\mathbf{C}} = m\left(\mathbf{\Sigma}\right) = \frac{1}{n}\sum_{i=1}^{n} u\left(\delta_{\boldsymbol{\mu},\mathbf{\Sigma}}(\boldsymbol{x}_i)\right)(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top. \tag{41}$$

Deriving (12) with respect to α leads to:

$$\alpha^* = \frac{\mathrm{tr}(\mathbf{\Sigma}) - \beta^* \mathbb{E}\left[\mathrm{tr}(\widetilde{\mathbf{C}})\right]}{p}. \tag{42}$$

Since the vectors $\boldsymbol{x}_i = \boldsymbol{\mu} + R_i \mathbf{\Lambda}\mathbf{U}_i$ for $1 \leq i \leq n$ are elliptically distributed and $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^\top$, then (see *e.g.* [26]):

$$(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top = R_i^2 \mathbf{\Lambda}\mathbf{U}_i\mathbf{U}_i^\top\mathbf{\Lambda}^\top$$

and $(\boldsymbol{x}_i-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})=R_i^2$. Thus,

$$\widetilde{\mathbf{C}}=\frac{1}{n}\sum_{i=1}^{n}u(R_i^2)R_i^2\boldsymbol{\Lambda}\mathbf{U}_i\mathbf{U}_i^\top\boldsymbol{\Lambda}^\top,$$

$$\text{and } \mathrm{tr}\left(\widetilde{\mathbf{C}}\right)=\frac{1}{n}\sum_{i=1}^{n}u(R_i^2)R_i^2\mathrm{tr}\left(\boldsymbol{\Lambda}\mathbf{U}_i\mathbf{U}_i^\top\boldsymbol{\Lambda}^\top\right) .$$

Using that R_i and $\mathbf{U}_i$ are independent and that $\mathbb{E}\left[\mathbf{U}_i\mathbf{U}_i^\top\right]=\frac{1}{p}\mathbf{I}$, it comes,

$$\begin{aligned}\mathbb{E}\left[\mathrm{tr}\left(\widetilde{\mathbf{C}}\right)\right]&=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[u(R_i^2)R_i^2\right]\mathrm{tr}\left(\boldsymbol{\Lambda}\mathbb{E}\left[\mathbf{U}_i\mathbf{U}_i^\top\right]\boldsymbol{\Lambda}^\top\right)\\&=f_1\mathrm{tr}(\boldsymbol{\Sigma})/p,\end{aligned}\tag{43}$$

where $f_1=\mathbb{E}\left[u(R^2)R^2\right]$ using R to represent the common distribution of the R_i's.

M-estimators satisfy the following relationship [6, 35]:

$$\mathbb{E}\left[u\left((\boldsymbol{x}-\boldsymbol{\mu})^\top(\sigma\boldsymbol{\Sigma})^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)(\sigma\boldsymbol{\Sigma})^{-1}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top\right]=\mathbf{I},$$

where σ is a positive scalar that depends on u and g (see *e.g.* [35] eq.(44) and (45)). Taking the trace on both sides and noticing that $\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top\right)=(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})=R^2$, it comes $\mathbb{E}\left[u(R^2/\sigma)R^2/\sigma\right]=p$. If $u(t)=-2g'(t)/g(t)$ then the quantity $f_1=\mathbb{E}\left[u(R^2)R^2\right]$ can be computed directly using integration by parts and is equal to p, which also shows that $\sigma=1$. For more general u, the value of f_1 is not straightforward. For $u(t)=-2g'(t)/g(t)$, the fact that $f_1=p$ leads thus to $\mathbb{E}\left[\mathrm{tr}(\widetilde{\mathbf{C}})\right]=\mathrm{tr}(\boldsymbol{\Sigma})$.

It follows (17), *i.e.* $\alpha^*=\frac{\mathrm{tr}(\boldsymbol{\Sigma})}{p}(1-\beta^*)$ that allows then to write $\widetilde{\boldsymbol{\Sigma}}_{\alpha\beta}=\beta\widetilde{\mathbf{C}}+(1-\beta)\frac{\mathrm{tr}(\boldsymbol{\Sigma})}{p}\mathbf{I}$ before deriving (12) w.r.t. β. This leads to:

$$\beta^*=\frac{\mathbb{E}\left[\mathrm{tr}\left(\widetilde{\mathbf{C}}\boldsymbol{\Sigma}\right)\right]-\frac{\mathrm{tr}(\boldsymbol{\Sigma})}{p}\mathbb{E}\left[\mathrm{tr}\left(\widetilde{\mathbf{C}}\right)\right]}{\mathbb{E}\left[\mathrm{tr}\left(\widetilde{\mathbf{C}}^2\right)\right]-2\frac{\mathrm{tr}(\boldsymbol{\Sigma})}{p}\mathbb{E}\left[\mathrm{tr}\left(\widetilde{\mathbf{C}}\right)\right]+\frac{\mathrm{tr}(\boldsymbol{\Sigma})^2}{p}}.\tag{44}$$

where $\mathbb{E}\left[\mathrm{tr}(\widetilde{\mathbf{C}})\right]=\mathrm{tr}(\boldsymbol{\Sigma})$ and similarly, $\mathbb{E}\left[\mathrm{tr}\left(\widetilde{\mathbf{C}}\boldsymbol{\Sigma}\right)\right]=f_1\mathrm{tr}\left(\boldsymbol{\Sigma}^2\right)/p=\mathrm{tr}(\boldsymbol{\Sigma}^2)$.

For the term $\mathbb{E}\left[\mathrm{tr}\left(\widetilde{\mathbf{C}}^2\right)\right]$, we can write:

$$\widetilde{\mathbf{C}}^2=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}u(R_i^2)R_i^2u(R_j^2)R_j^2\boldsymbol{\Lambda}\mathbf{U}_i\mathbf{U}_i^\top\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda}\mathbf{U}_j\mathbf{U}_j^\top\boldsymbol{\Lambda}^\top,$$

and using the mutual independence of all $\mathbf{U}_i$'s and R_i's and their common distribution represented by variable $\mathbf{U}$ and R respectively, it comes,

$$\begin{aligned}\mathbb{E}\left[\mathrm{tr}(\widetilde{\mathbf{C}}^2)\right] &= \frac{1}{n^2}\Big(\sum_{i\neq j}\mathbb{E}\left[u(R^2)R^2\right]^2\mathrm{tr}\left((\mathbf{\Lambda}\mathbb{E}\left[\mathbf{U}\mathbf{U}^\top\right]\mathbf{\Lambda}^\top)^2\right)\\ &\quad+\sum_{i=1}^{n}\mathbb{E}\left[(u(R^2)R^2)^2\right]\mathbb{E}\left[\mathrm{tr}\left((\mathbf{U}^\top\mathbf{\Lambda}^\top\mathbf{\Lambda}\mathbf{U})^2\right)\right]\Big)\\ &= \frac{n-1}{n}\frac{f_1^2\,\mathrm{tr}(\mathbf{\Sigma}^2)}{p^2}+\frac{1}{n}f_2\,\mathbb{E}\left[(\mathbf{U}^\top\mathbf{\Lambda}^\top\mathbf{\Lambda}\mathbf{U})^2\right]\ .\end{aligned}$$

Following [8], to compute the last expectation, we can use the fact that for an elliptical distribution, the particular decomposition of $\mathbf{\Sigma}$ is irrelevant. In particular we can take $\mathbf{\Lambda}=\mathbf{D}\mathbf{A}^{1/2}$ where $\mathbf{\Sigma}=\mathbf{D}\mathbf{A}\mathbf{D}^T$ is the eigenvalue decomposition of $\mathbf{\Sigma}$. Thus $\mathbf{\Lambda}^\top\mathbf{\Lambda}=\mathbf{A}$ where $\mathbf{A}$ is the diagonal matrix containing the eigenvalues of $\mathbf{\Sigma}$. For our purpose, we actually only need that $\mathbf{A}=\mathbf{\Lambda}^T\mathbf{\Lambda}$ is diagonal since then $\mathrm{tr}(\mathbf{A})=\mathrm{tr}(\mathbf{\Sigma})$ and $\mathrm{tr}(\mathbf{A}^2)=\mathrm{tr}(\mathbf{\Sigma}^2)$. Let $a_1,\ldots,a_p$ be the diagonal entries of $\mathbf{A}$ and denote the elements of $\mathbf{U}$ as $\mathbf{U}=(u_1,\ldots,u_p)^\top$, the last expectation can be computed as,

$$\begin{aligned}\mathbb{E}\left[\left(\mathbf{U}^\top\mathbf{A}\mathbf{U}\right)^2\right] &= \mathbb{E}\left[\sum_{k=1}^{p}a_k^2u_k^4+\sum_{k\neq m}a_ka_mu_k^2u_m^2\right]\\ &= \sum_{k=1}^{p}a_k^2\mathbb{E}\left[u_k^4\right]+\sum_{k\neq m}a_ka_m\mathbb{E}\left[u_k^2u_m^2\right].\end{aligned}$$

At last, it comes from Lemma 2 in [51] that , $\mathbb{E}\left[u_k^4\right]=\frac{3}{p(p+2)}$ and $\mathbb{E}\left[u_k^2u_m^2\right]=\frac{1}{p(p+2)}$ for $k\neq m$. Then,

$$\begin{aligned}\mathbb{E}\left[\left(\mathbf{U}^\top\mathbf{A}\mathbf{U}\right)^2\right] &= \frac{1}{p(p+2)}\left(3\sum_{k=1}^{p}a_k^2+\sum_{k\neq m}a_ka_m\right)\\ &= \frac{1}{p(p+2)}\left(2\mathrm{tr}(\mathbf{A}^2)+(\mathrm{tr}(\mathbf{A}))^2\right)\\ &= \frac{1}{p(p+2)}\left(2\mathrm{tr}(\mathbf{\Sigma}^2)+\mathrm{tr}(\mathbf{\Sigma})^2\right).\end{aligned}$$

We finally get:

$$\begin{aligned}\mathbb{E}\left[\mathrm{tr}(\widetilde{\mathbf{C}}^2)\right] =& \frac{1}{n^2}\left[n\left(f_2\frac{1}{p(p+2)}\left(2\mathrm{tr}(\mathbf{\Sigma}^2)+\mathrm{tr}(\mathbf{\Sigma})^2\right)\right)\right.\\ &\left.+(n^2-n)\left(f_1^2\frac{1}{p^2}\mathrm{tr}(\mathbf{\Sigma}^2)\right)\right]\\ =&\mathrm{tr}(\mathbf{\Sigma}^2)\left(\frac{n-1}{n}+f_2\frac{2}{np(p+2)}\right)\\ &+\mathrm{tr}(\mathbf{\Sigma})^2 f_2\frac{1}{np(p+2)}.\end{aligned}$$

The shrinkage coefficient β^* thus takes the form:

$$\beta^* = \frac{\mathrm{tr}(\mathbf{\Sigma}^2)-\frac{\mathrm{tr}(\mathbf{\Sigma})^2}{p}}{\mathrm{tr}(\mathbf{\Sigma}^2)\left(\frac{n-1}{n}+\frac{2f_2}{np(p+2)}\right)+\frac{f_2\mathrm{tr}(\mathbf{\Sigma})^2}{np(p+2)}-\frac{\mathrm{tr}(\mathbf{\Sigma})^2}{p}},$$

and the result is proved.

Note also that this expression of β^* is by construction always in $[0,1]$ if $p \geq 2$. Indeed, the difference between the numerator and denominator is

$$\mathrm{tr}(\mathbf{\Sigma}^2)\left(\frac{2f_2}{np(p+2)}-\frac{1}{n}\right)+\frac{f_2}{np(p+2)}\mathrm{tr}(\mathbf{\Sigma})^2.$$

Since $f_2 = \mathbb{E}\left[(u(R^2)R^2)^2\right]$, Jensen inequality gives $f_2 \geq f_1^2 = p^2$, and

$$\frac{2f_2}{np(p+2)}-\frac{1}{n} \geq \frac{p-2}{n(p+2)} \geq 0,$$

and $\beta^* \leq 1$. Similarly, since $\mathbf{\Sigma}$ is a symmetric matrix, we have $\mathrm{tr}(\mathbf{\Sigma})^2 \leq \mathrm{tr}(\mathbf{\Sigma}^2)p$, and therefore $\beta^* \geq 0$. □

8.2. Proof of Proposition 3.1

Let us denote f_R the pdf of R. According to Theorem 2.9 in [28], we have:

$$f_R(t) = \frac{(2\pi)^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2}\right)}t^{p-1}g(t^2) \Leftrightarrow g(t) = \frac{\Gamma\left(\frac{p}{2}\right)}{(2\pi)^{\frac{p}{2}}}t^{(1-p)/2}f_R(\sqrt{t}).$$

We prove each case separately.

1. If R^2 is a Chi-squared distribution with p degrees of freedom, then:

$$g(t) = \frac{\Gamma\left(\frac{p}{2}\right)}{(2\pi)^{\frac{p}{2}}}t^{(1-p)/2}f_R(\sqrt{t}) = (2\pi)^{-p/2}\exp\left(-\frac{t}{2}\right).$$

It thus comes $u(t) = 1$, and $f_2 = \mathbb{E}\left[\chi_p^4\right]$. The moments of the Chi-squared distribution lead to $f_2 = p(p+2)$.

2. Some straightforward calculations lead to:

$$g(t) = \frac{\Gamma\left(\frac{p}{2}\right)}{\Gamma\left(\frac{2q+p-2}{2s}\right)} \frac{\lambda^{\frac{2q+p-2}{2s}}}{\pi^{\frac{p}{2}}} t^{q-1} \exp\left(-\lambda t^s\right)$$

and therefore

$$u(t) = 2\frac{1-q}{t} + 2\lambda s t^{s-1}.$$

By noticing that $u(t)^2 t^2 = 4(1-q)^2 + 8\lambda s(1-q)t^s + 4\lambda^2 s^2 t^{2s}$, the moment $\mathbb{E}\left[u(R^2)^2 R^4\right]$ may thus be rewritten as follows:

$$4(1-q)^2 + 4s(1-q)\mathbb{E}\left[\chi^2_{\frac{2q+p-2}{s}}\right] + s^2\mathbb{E}\left[\left(\chi^2_{\frac{2q+p-2}{s}}\right)^2\right].$$

Using the moments of the Chi-squared distribution concludes the proof.

3. If R^2/p is a Fisher distribution with p and ν degrees of freedom, it is known that $\boldsymbol{x}$ follows a p-variate t-distribution with ν degrees of freedom. In this case, $g(t)$ is given by $(\nu\pi)^{-p/2}\left(1 + t/\nu\right)^{-(p+\nu)/2}$, and therefore $u(t) = (p+\nu)/(t+\nu)$. In addition,

$$\begin{aligned}
\mathbb{E}\left[u(R^2)^2 R^4\right] &= \int_0^\infty \left(u(r^2)\right)^2 r^4 f_R(r) dr \\
&= \int_0^\infty 4\frac{\left(g'(r^2)\right)^2}{\left(g(r^2)\right)^2} r^4 \frac{2\pi^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2}\right)} r^{p-1} g(r^2) dr \\
&= \int_0^\infty 8\frac{\left(g'(r^2)\right)^2}{g(r^2)} \frac{\pi^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2}\right)} r^{p+3} dr \\
&= 8\frac{\pi^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2}\right)} \int_0^\infty \frac{\left(g'(r^2)\right)^2}{g(r^2)} r^{p+3} dr.
\end{aligned}$$

The ratio $(g'(r^2))^2/g(r^2)$ may be rewritten here:

$$\frac{\left(g'_{r^2}(r^2)\right)^2}{g(r^2)} = \frac{1}{4}\frac{\Gamma\left(\frac{p+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\pi^{\frac{p}{2}}} \nu^{-\frac{p}{2}-2}(\nu+p)^2 \left(1 + \frac{r^2}{\nu}\right)^{-\frac{\nu+p+4}{2}}.$$

Combining the previous relationships obtained, it comes:

$$\begin{aligned}
\mathbb{E}\left[u(R^2)^2 R^4\right] = &\frac{\Gamma\left(\frac{p+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\Gamma\left(\frac{p}{2}\right)} (\nu+p)^2 \\
&\times \int_0^\infty \left(\frac{r^2}{\nu}\right)^{\frac{p}{2}+1} \left(1 + \frac{r^2}{\nu}\right)^{-\frac{\nu+p}{2}-2} d\left(\frac{r^2}{\nu}\right).
\end{aligned}$$

Using the known moments of the t-distribution leads to:

$$\begin{aligned}\mathbb{E}\left[u(R^2)^2 R^4\right] &= \frac{\Gamma\left(\frac{p+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\Gamma\left(\frac{p}{2}\right)}(\nu+p)^2\frac{\Gamma\left(\frac{p}{2}+2\right)\Gamma\left(\frac{\nu}{2}\right)}{\Gamma\left(\frac{p+\nu}{2}+2\right)} \\ &= \frac{(\nu+p)(p+2)p}{p+\nu+2}.\end{aligned}$$

□

Acknowledgments

Authors acknowledge LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) and the Grenoble Alpes Data Institute supported by the French National Research Agency under the "Investissements d'avenir" program (ANR-15-IDEX-02).

References

[1] O. Ledoit, M. Wolf, Improved estimation of the covariance matrix of stock returns with an application to portfolio selection, Journal of empirical finance 10 (5) (2003) 603–621.

[2] J. H. Friedman, Regularized discriminant analysis, Journal of the American statistical association 84 (405) (1989) 165–175.

[3] M. Finegold, M. Drton, Robust graphical modeling of gene networks using classical and alternative t-distributions, The Annals of Applied Statistics 5 (2A) (2011) 1057–1080.

[4] J. Owen, R. Rabinovitch, On the class of elliptical distributions and their applications to the theory of portfolio choice, The Journal of Finance 38 (3) (1983) 745–752.

[5] T. Bodnar, W. Schmid, A test for the weights of the global minimum variance portfolio in an elliptical model, Metrika 67 (2) (2008) 127.

[6] R. A. Maronna, Robust M-Estimators of Multivariate Location and Scatter, The Annals of Statistics 4 (1) (1976) 51–67.

[7] M. Chen, C. Gao, Z. Ren, Robust Covariance and Scatter Matrix Estimation under Huber's Contamination Model, The Annals of Statistics 46 (2018) 1932–1960.

[8] Y. Chen, A. Wiesel, A. O. Hero, Robust Shrinkage Estimation of High-Dimensional Covariance Matrices, IEEE Transactions on Signal Processing 59 (9) (2011) 4097–4107.

[9] M. Yuan, Y. Lin, Model selection and estimation in the Gaussian graphical model, Biometrika 94 (2007) 19–35.

[10] O. Banerjee, L. El Ghaoui, A. d'Aspremont, Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data, Journal of Machine learning research 9 (Mar) (2008) 485–516.

[11] J. Friedman, T. Hastie, R. Tibshirani, Sparse inverse covariance estimation with the graphical Lasso, Biostatistics 9 (3) (2008) 432–441.

[12] N. Meinshausen, P. Buhlmann, High-dimensional graphs and variable selection with the Lasso, The Annals of Statistics 34 (3) (2006) 1436–1462.

[13] C. Giraud, Estimation of Gaussian graphs by model selection, Electronic Journal of Statistics 2 (2008) 542–563.

[14] J. Jankova, S. van de Geer, Inference in high-dimensional graphical models (2018). arXiv:1801.08512
URL https://arxiv.org/abs/1801.08512

[15] K. Ashurbekova, S. Achard, F. Forbes, Robust structure learning using multivariate T-distributions, in: 50e Journees de la Statistique (JdS'2018), Saclay, France, 2018.

[16] H. Liu, F. Han, M. Yuan, J. Lafferty, L. Wasserman, High-dimensional semiparametric Gaussian copula graphical models, The Annals of Statistics 40 (4) (2012) 2293–2326.

[17] T. Zhao, H. Liu, Calibrated precision matrix estimation for high-dimensional elliptical distributions, IEEE transactions on information theory 60 (2014) 7874–7887.

[18] O. Ledoit, M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, Journal of Multivariate Analysis 88 (2) (2004) 365–411.

[19] D. E. Tyler, A Distribution-free M-estimator of Multivariate Scatter, The Annals of Statistics 15 (1) (1987) 234–251.

[20] F. Pascal, Y. Chitour, J. Ovarlez, P. Forster, P. Larzabal, Covariance Structure Maximum-Likelihood Estimates in Compound Gaussian Noise: Existence and Algorithm Analysis, IEEE Transactions on Signal Processing 56 (1) (2008) 34–48.

[21] E. Ollila, H. Oja, V. Koivunen, Complex-valued ICA based on a pair of generalized covariance matrices, Computational Statistics & Data Analysis 52 (2008) 3789–3805.

[22] A. Wiesel, Geodesic convexity and covariance estimation, IEEE Transactions on Signal Processing 60 (2012) 6182–6189.

[23] Y. Sun, P. Babu, D. P. Palomar, Regularized robust estimation of mean and covariance matrix under heavy-tailed distributions, IEEE Transactions on Signal Processing 63 (12) (2015) 3096–3109.

[24] J. Liu, D. Palomar, Regularized robust estimation of mean and covariance matrix for incomplete data, Signal Processing 165 (2019) 278 – 291.

[25] J. Tong, R. Hu, J. Xi, Z. Xiao, Q. Guo, Y. Yu, Linear shrinkage estimation of covariance matrices using low-complexity cross-validation, Signal Processing 148 (C) (2018) 223–233.

[26] S. Cambanis, S. Huang, G. Simons, On the theory of elliptically contoured distributions, Journal of Multivariate Analysis 11 (1981) 368–385.

[27] D. Kelker, Distribution theory of spherical distributions and a location-scale parameter generalization, Sankhya: The Indian Journal of Statistics, Series A 32 (4) (1970) 419–430.

[28] K.-T. Fang, S. Kotz, K. W. Ng, Symmetric multivariate and related distributions, Chapman and Hall, 1990.

[29] G. Frahm, Generalized elliptical distributions: Theory and applications, Ph.D. thesis, Universität zu Köln (2004).

[30] H. J. Malik, B. Abraham, et al., Multivariate logistic distributions, The Annals of Statistics 1 (3) (1973) 588–590.

[31] S. Nadarajah, The Kotz-type distribution with applications, Statistics: A Journal of Theoretical and Applied Statistics 37 (4) (2003) 341–358.

[32] T. Eltoft, T. Kim, T.-W. Lee, On the multivariate Laplace distribution, IEEE Signal Processing Letters 13 (5) (2006) 300–303.

[33] O. Arslan, A. I. Genç, A generalization of the multivariate slash distribution, Journal of Statistical Planning and Inference 139 (3) (2009) 1164–1170.

[34] Y. Zhao, Covariance matrices of quadratic forms in elliptical distributions, Statistics & Probability Letters 21 (2) (1994) 131–140.

[35] E. Ollila, D. E. Tyler, V. Koivunen, H. V. Poor, Complex elliptically symmetric distributions: Survey, new results and applications, IEEE Transactions on Signal Processing 60 (11) (2012) 5597–5625.

[36] E. Ollila, D. E. Tyler, Regularized M-estimators of scatter matrix, IEEE Transactions on Signal Processing 62 (22) (2014) 6059–6070.

[37] D. E. Tyler, Statistical analysis for the angular central Gaussian distribution on the sphere, Biometrika 74 (3) (1987) 579–589.

[38] R. J. Muirhead, Aspects of multivariate statistical theory, Vol. 197, John Wiley & Sons, 2009.

[39] P. J. Huber, Robust estimation of a location parameter, The Annals of Mathematical Statistics 35 (1) (1964) 73–101.

[40] T. Zhang, A. Wiesel, Automatic diagonal loading for Tyler's robust covariance estimator, in: 2016 IEEE Statistical Signal Processing Workshop (SSP), IEEE, 2016, pp. 1–5.

[41] E. Hashorva, On the regular variation of elliptical random vectors, Statistics & probability letters 76 (14) (2006) 1427–1434.

[42] B. M. Hill, A simple general approach to inference about the tail of a distribution, The Annals of Statistics 3 (5) (1975) 1163–1174.

[43] L. de Haan, A. Ferreira, Extreme value theory: an introduction, Springer Science & Business Media, 2006.

[44] A. Usseglio-Carleve, Estimation of conditional extreme risk measures from heavy-tailed elliptical random vectors, Electronic Journal of Statistics 12 (2018) 4057–4093.

[45] E. Gómez-Sánchez-Manzano, M. Gómez-Villegas, J. Marín, Sequences of elliptical distributions and mixtures of normal distributions, Journal of Multivariate Analysis 97 (2) (2006) 295–310.

[46] M. Termenon, C. Delon-Martin, A. Jaillard, S. Achard, Reliability of graph analysis of resting state fMRI using test-retest dataset from the human connectome project, Neuroimage 142 (15) (2016) 172–187.

[47] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, M. Joliot, Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain, Neuroimage 15 (1) (2002) 273–289.

[48] H. Chen, Y. Xia, A nonparametric normality test for high-dimensional data (2019). arXiv:1904.05289

[49] C. Liu, D. B. Rubin, ML estimation of the t distribution using EM and its extensions, ECM and ECME, Statistica Sinica (1995) 19–39.

[50] M. Rubinov, O. Sporns, Complex network measures of brain connectivity: uses and interpretations, Neuroimage 52 (3) (2010) 1059–1069.

[51] Y. Chen, A. Wiesel, Y. C. Eldar, A. O. Hero, Shrinkage algorithms for MMSE covariance estimation, IEEE Transactions on Signal Processing 58 (10) (2010) 5016–5029.